DOCUMENT RESUME

ED 115 858 CE 005 696

AUTHOR	Shirley, Maurice
TITLE	Computer Specifications for the MISOE [Management Information System for Occupational Education] Prototype System.
INSTITUTION	Management and Information System for Occupational Education, Winchester, Mass.
SPONS AGENCY	Massachusetts State Dept. of Education, Boston. Div. of Occupational Education.
NOTE	63p.; For related documents, see ED 062 553; ED 068 646-647; ED 072 225; ED 072 228; ED 072 303-304; CE 005 687-727
EDRS PRICE	MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS	*Computer Programs; Computers; Computer Science; Demonstration Projects; Electronic Data Processing; *Estimated Costs; *Management Information Systems; Policy Formation; *Specifications; State Programs; *Vocational Education
IDENTIFIERS	Collaborative MISOE; Management Information System Occupational Educa; MISOE

ABSTRACT

The report was prepared by a member of the staff of the Management Information System for Occupational Education (MISOE). It constitutes a clear and precise statement of the tasks required to establish a "laboratory MISOE," as described in the Journal of Research and Development in Education (Winter 1974), to accomplish the purposes of "Collaborative MISOE." Collaborative MISOE's major goal is to provide a "real time" experience for MISOE developers and occupational education policy-makers such that future development of interactive computer systems designed to facilitate improved policy can be based on human experience. The report specifies clearly the alternative required to accomplish this goal. MISOE was a major development effort to improve the policy-making function at the State level, undertaken in Massachusetts in 1971 and ended in 1975. MISOE sought to combine three large data sets (a census data system, a longitudinal sample data system, and an ex post facto data set) on a single computer with interactive capability, bringing to the policy-makers' fingertips a range of psychological, economic, and demographic data and providing them with a knowledge base for "parametizing" forecast models. Proposals for the adaptation of existing systems to MISOE's needs and discussion of these proposals is included. (Author/AJ)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED115858

COMPUTER SPECIFICATIONS FOR THE MISOE PROTOTYPE SYSTEM

CE005698

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Maurice Shirley
Management Information System for Occupational Education

Table of Contents

Page

Introduction. 2

MISOE "State 51" Data Files 3

MISOE Interactive System 7

Data Analysis 8

Terminals 8

Software Systems 9

Specifications for MISOE "State 51" Feasibility Study 10

Supplemental Specification 13

Dr. John J. Donovan
Associate Professor of Management Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Proposals 16

The Consistent System 16

XSIM 28

Response to Proposal (1) 30

Dr. John J. Donovan
Associate Professor of Management Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Response to Proposal (2) 33

Maurice Shirley
MISOE Staff

Cost Estimates for Software 37

Terminal Configurations 41

Recommendations 44

Table of Contents (Continued) Page

Appendix I

Analysis Capabilities of The Consistent and XSIM Systems for MISOE 47

Dr. John A. Creager
Division Director
Division of Educational Statistics
American Council on Education
Washington, D. C.

Appendix II

Cost Projections to Real MISOE 57

Tables

1 - Approximate Size of "State 51" Data Files 6

2 - Prototype System Cost Estimate 40

3 - Terminal Configuration Costs 43

4 - Estimated Annual Data Processing Costs 59

Preface

This report was prepared by Maurice Shirley of the MISOE staff. It constitutes a clear and precise statement of the tasks required to establish a "laboratory MISOE" as described in the Journal of Research and Development in Education, Winter, 1974, Vol. 7, No. 2 to accomplish the purposes of "Collaborative MISOE" (see Multi-State Proposals for "Collaborative MISOE"). Fundamentally, the major goal of "Collaborative MISOE" is to provide a "real time" experience for MISOE developers and Occupational Education policy makers such that future development of interactive computer systems designed to facilitate improved policy can be based on human experience. This report specifies clearly the alternative required to accomplish this goal.

Mr. Shirley was assisted in this report by Dr. John A. Creager, Division Director, Division of Educational Statistics, American Council on Education, Washington, D. C., and Mr. Michael Garet, Research Assistant, MISOE staff.

Dr. William G. Conroy, Jr.
Lowell, Massachusetts

Introduction

The MISOE Prototype System (State '51) is designed to demonstrate and test the interactive capabilities of MISOE by allowing potential users to formulate and evaluate policies on the basis of data in the system information files. Although the system will contain "dummy" rather than real data, the data will be comparable in volume and scope to what could be expected if real data were collected for any state participating in MISOE. The MISOE Prototype System will allow prospective users to explore the full capabilities of MISOE as described in the Journal of Research and Development in Education, Winter, 1974, Vol. 7, No. 2.

This report begins with a description of the MISOE data and a discussion of the interactive system data management, report generating and data analysis requirements. Proposals by the developers of several data management packages follow, describing how their systems operate and offering cost estimates for implementing the MISOE Prototype System.

Next are responses on these proposed systems by Dr. John J. Donovan, Associate Professor of Management Science at M.I.T., and Maurice Shirley of the MISOE staff. The final section of the report consists of analysis and cost estimates for the proposed systems prepared by the MISOE data processing staff and recommendations for implementation. Appendix I discusses the data analysis capability of the systems under consideration by Dr. John A. Creager, Division Director, Division of Educational Statistics, American Council on Education, while Appendix II examines cost projections from "State 51" to "real MISOE".

MISOE "State 51" Data Files

The two major sections of the MISOE data files are the Census Data System (CDS) and the Sample Data System (SDS). The Census Data System (CDS) has basic information (expenditures, enrollment, staff and instructional area utilization) by OE program describing all OE programs in the state. CDS has been designed to be integrated with the Massachusetts Department of Education census data system such that basic data describing both occupational and non-occupational education can be reconciled from the school to the state level. The Sample Data System (SDS) has detailed information describing Occupational Education programs, including student characteristics as they enter these programs, the educational process, the product or outcomes of this process and the impact of the educational process for discreet groups of students on society and self over time. A more detailed description of CDS and SDS can be found in Chapter 3 and 4, respectively, of the Journal of Research and Development in Education, Vol. 7, No. 2, 1974.

The prototype SDS files will be developed in "State 51" described in Chapter 4 of the Journal including the INPUT, PROCESS, PRODUCT AND IMPACT batteries as well as basic identification data for each student, teacher or administrator in the sample. The SDS files will also contain cost data for each program, on a per student basis, broken down by line item for occupational and nonoccupational education, as well as enrollment data for each program by grade and level, including attendance and dropout status. The unit of analysis for SDS is the individual student.

The MISOE prototype "State 51" SDS files will have data on 15,000 students, 1,000 teachers, 200 administrators and 500 educational programs. The amount of data for 1 cohort (including updates and follow-up studies) of is approximately 30 million characters.

The CDS files used in estimating costs for the prototype system are a truncated version of the final CDS files. The reporting forms for CDS are being revised as this report is being prepared, making it not possible to generate an exact prototype file, but differences between "State 51" and operational MISOE will be slight.

The prototype CDS files will contain data describing:

(1) Costs by line item, by OE/NOE programs for schools and LEAs at the secondary and post-secondary levels.

(2) Enrollment by grade, sex and program at the school, LEA and secondary-post-secondary levels.

(3) Enrollment by student group within OE programs (see Journal), by type, race and sex. The student groups (GID) is the unit of analysis in CDS.

(4) Description of administrative staff at all levels described in (1) and (2).

(5) Descriptions of specific teacher characteristics at all levels described in (1) and (2).

(6) Teacher assignments by GID for all levels described in (1) and (2).

(7) Instructional area utilization by OE programs within schools.

(8) Student class time by GID within OE programs.

(9) TERMOBS by GID within OE program.

(10) Attrition for OE (by GID) and nonoccupational at all levels described in (1) and (2), above.

The "State 51" CDS files will contain data on 250 LEAs, 350 schools 6,000 teachers and 3,500 programs. The amount of data for each year of the truncated version of CDS is approximately 10 million characters.

Table I illustrates the size of the prototype system data files (as used for cost estimates) by listing the various SDS and CDS (truncated) data files and the number of records expected in each file along with the number of items in each record.

TABLE 1

Approximate Size of "State 51" Data Files

FILES	NUMBER OF RECORDS	NUMBER OF ITEMS EACH RECORD
SDS		
Student Basic Identification	15,000.	156.
Student Input Battery 1	10,825.	391.
Student Input Battery 2	4,175.	388.
Student Process	6,540.	32.
Student Product	15,000.	17.
Student TERMOBS	6,540.	140.
Student Impact	5,000.	801.
Teacher Basic Identification	1,000.	40.
Teacher Process	1,000.	521.
Administrator Basic Identification	200.	40.
Administrator Process	200.	285.
SDS Program	500.	71.
CDS (Truncated)		
CDS Program	3,500	526
CDS School	350	420
CDS Teachers	6,000	41
CDS LEA	250	65

MISOE INTERACTIVE SYSTEM

The performance goals that the MISOE interactive system will try to achieve are described in this section. These are things that we feel the system should be able to do to operate effectively in supporting the policy making process.

To allow the user to interact freely with the system, MISOE requires extremely flexible data management. The instructions issued by the user must be quite simple and correspond closely to language statements (e.g., to calculate the mean, the command should use the word "mean"). The command language should make it easy for users who are not experienced programmers to manipulate, retrieve and analyze data from the MISOE files with minimum training.

The system should prompt the user to supply missing parameters if his instructions are not complete. Output reports should appear at the terminal in standard format but with provisions for the user to issue explicit instructions when desired to control the spacings of columns, rows and headers and to request graphs or off-line printouts of lengthy reports.

The system should be able to key on any field in the data files and to use any field in any file in a conditional statement. The user must be able to create sub-files from the original data and store them either temporarily or permanently. The system should catalog and store routines or mathematical formulae that the user will use repeatedly. While the interactive system should allow data to be typed in at a terminal, it is expected that the bulk of the data will be loaded off-line from punched cards or tape. The system should have provisions to edit, correct or modify data after loading.

There should be security provisions to allow certain users (MISOE staff) to read and write data files while other users (managers) can read files, but not alter them. Managers should be able to create and store sub-files for personal use, but the updating of any original file should be a function of the MISOE staff. An accounting system should be a part of the system, recording not only connect and CPU time by user, but also the names of files accessed and the cataloged routines used.

Data Analysis

In addition to supporting ordinary data analysis (totals, counts, means), the MISOE system places emphasis on statistical measures of relationship (correlation, regression, etc.). Table 3, page 81 of the Journal lists statistical routines that should be available in the system. It should be possible to access data directly with these routines (i.e., preparation of intermediate files for access by a separate statistical package should be unnecessary.) For the MISOE Sample Data System (SDS), the statistical routines must consider the sample weights attached to each data record. The CDS data will, of course, be unweighted. The interactive system should also provide a programming language appropriate for simulating dynamic models, thus allowing users to investigate the consequences over time of alternate occupational education policies through the calculus of Dynamic Simulation, Mathematical Programming and Path Analysis.

Terminals

It is anticipated that the MISOE Prototype System will be used as a "learning laboratory" for management and MISOE staff, often with groups of people observing the interaction between a terminal operator and the data

files. For this a CRT terminal with a graphics capability and one or more large screen monitors would be ideal. As part of the "learning laboratory" participants will be divided into small groups for workshops, with each group operating its own terminal. For this MISOE will use less expensive typewriter terminals.

MISOE also needs to make printed copies of the reports generated by the computer system. Storing and editing data off-line on either tapes or disks would be a very desirable feature of the terminal system, as much of the "State 51" data could be prepared at the terminal rather than keypunched. Other desirable features of a terminal are: (1) reasonably quiet and dependable in operation, and, (2) a transmission speed of at least 30 characters per second.

Software Systems

Three software systems having data management and statistical analysis capabilities were judged provisionally suitable for the MISOE interactive system. The developers of these systems were requested to submit reports describing how and at what cost their systems could be used for MISOE "State 51".

The three developers were:

- National CSS, Inc. - using an interactive version of RAMIS on their time-sharing system.
- Dynamic Associates, Inc. - developers of XSIM to be used on Interactive Data Corporation's time-sharing system.
- The Cambridge Project, M.I.T. - the developers of the Consistent System with the JANUS data management system to be used on M.I.T.'s MULTICS System.

National CSS, Inc., after studying MISOE specifications, determined that their interactive version of RAMIS could not meet the needs of MISOE unless it was considerably revised at an exorbitant cost. Therefore, we did not consider using RAMIS for the MISOE prototype system.

The Cambridge Project and Dynamics Associates, Inc. both felt their systems were well suited to MISOE's needs and their reports follow the specifications on which their reports were based.

Specifications for MISOE "State 51" Feasibility Study

1. MISOE "State 51" will be a prototype system installed on a commercial time-sharing facility using a data management package having adequate statistical and report generating capabilities. It will be complete with "dummy" data and will be used to study the usefulness of interactive capabilities of the MISOE Census Data System (CDS) and the MISOE Sample Data System (SDS) as they are described in the Journal of Research and Development in Education, Vol. 7, No. 2, 1974. The prototype system must be capable of expansion at a later date to accommodate several states each having their own CDS and SDS data bases.

2. The data management system will:

a) Operate interactively in response to requests by the user to manipulate, retrieve and analyze specific data as described in Chapter 8 of the Journal.

b) Generate tabular and graphic reports (scatter plots, bar graphs, time plots and pie charts). On-line tables and graphs must be displayable on a CRT type terminal (the proposal should specify terminals that can be used). Lengthy tabular reports should be printed off-line.

c) Accommodate input data from magnetic tapes or punch cards (both by batch process) or from a terminal.

d) Generate (test) data internally.

e) Create new variables and/or new files and store them either temporarily or permanently.

f) Update data files.

g) Key on any field in the data files.

h) Provide for separate backup storage (on tape, for example) of both CDS and SDS data bases.

I) Allow MISOE to store the entire system off-line when not in use.

j) Provide routines so that all data accessed from the Sample Data System (SDS) will be weighted to reflect the aggregate populations.

3. User commands (at level I as defined by Dr. Donovan, below) should be as simple and easy to use as possible and make use of an INDEX as described in Chapter 7 of the Journal.

4. The MISOE System should be capable of Dynamic Simulation (this could be a separate package but must obtain data from the CDS and SDS files).

5. Size of MISOE "State 51":

a) CDS files will have 350 schools and 3500 educational programs.

b) SDS files will have 15,000 students, 1,000 teachers and 500 educational programs.

c) SDS files will be for 1 cohort (1 class) with updates and follow-up studies.

11

d) CDS files will be for one year.

e) An accompanying folder contains detailed record layouts for all of the files in MISOE-CDS and SDS.

6. The MISOE "State 51" System is to be designed, installed and loaded with "dummy" data within a period of two months.

7. "State 51" will be in operation for demonstration purposes for about eight weeks.

8. "State 51" cost estimates should include:

a) Cost to install the initial system.

b) Cost to load data from tape files (estimate a unit price).

c) Cost to generate test data internally (unit price).

d) Cost to allocate new permanent files (unit price).

e) Storage costs during the entire life of "State 51" (3 months to install and debug plus about 8 weeks operation).

f) Cost to reload the entire system after off-line storage.

g) Operating costs - estimates of costs to statistically analyze data and generate specific reports chosen from Chapter 8 of the Journal--we suggest costing out at least five tables chosen from Tables 1, 3, 4, 5, 8, 9, 10, 11, 22 and 25 in Chapter 8.

h) The Journal of Research and Development in Education, Vol. 7, No. 2, 1974 and the supplemental specifications of Dr. Donovan (which follows), should both be considered part of this specification.

Supplemental Specifications
Professor John J. Donovan, M.I.T.

This statement is to supplement the specifications of the MISOE System as described in the Journal of Research and Development in Education, Volume 7, No. 2, 1974. To create a framework for implementation, we define five levels of distinct users. The motivation for these distinctions is to clear up the confusion of needs and responsibilities of different users. For example, on performance information, what type of user needs to know about the responsiveness of the system, cost of queries, or information as to the number of I/O seeks on a disk?

The five levels are:

1) Policy makers.

These are the ultimate users (e.g., people in the State Department of Education). There are two classifications of users at this level--the user making policy decisions and the people that are entering data into the system.

2) System customizers.

These are the people that are responsible for creating new files, creating new file structures, creating new entries in existing files, creating new commands, making abbreviations of old commands and supplying new statistical packages. It is envisioned that this level of user will correspond to the MISOE programming staff.

3) Information systems programmers.

These are the people that are responsible for building

the tools for accessing and storing information in the files. We expect to contract this responsibility out to a proprietary group such as RAMIS. These people are responsible for the initial system. The initial system and the tools for creating new files will be used by the system customizer (level 2 user) in the future to adapt the system to the evolving needs of level 1 users.

4) Systems maintenance users.

These are the people that are responsible for the software that was used to implement the tools of level 3, for example, the compilers and the operating systems. If XSIM is used, for example, then level 4 corresponds to the people that are concerned with maintaining the language AED.

5) Hardware.

These people are concerned with building, maintaining and operating the hardware of the system. We expect this level of user will be a manufacturer.

Functions and Facilities of Each Level:

Level 1

The Journal, in general, describes the facilities that a user needs at level 1, the types of questions he will ask, the types of statistics he needs, e.g., cost of certain educational programs. Example of additional facilities are computer programs for validation of the data, batch facilities for entering large amounts of data, facilities for production of specialized publications and reports, etc.

Level 2

A static description of the files which a level 2 user requires is presented in the Journal. However, there are some additional facilities he needs to respond to the dynamic and evolutionary nature of the system needs of the level 1 user. For example, he needs performance measurements as to the cost and the responsiveness of facilities that the user of level 1 needs.

We can say with certainty that at this point in time we have not anticipated all the types of questions that a level 1 user will ask of the system, nor do we know all of the types of data we will need or want to keep. Thus, a level 2 user also needs tools for entering new fields, that is, data that was not originally contemplated being in the system. He needs facilities for keying on data in the system in ways that were originally not anticipated.

Level 3

The level 3 user, which we assume will be an outside contractor, will be concerned with the implementation of tools and a file system that can meet the needs of level 1 and level 2 users. As such, he is concerned with the reliability of the system, size of files, host language, performance measurements, host operating system and, most importantly, production of flexible tools to allow the system customizers to adjust the system to meet the evolutionary needs of MISOE.

Level 4

The level 4 user will be (it is anticipated) an outside manufacturer and will be concerned with the size of the second storage device, compiler maintenance and system utilization.

Level 5

The level 5 user will be a manufacturer and he is concerned with hardware.

PROPOSALS

The following are two proposals by data management policy developers for adapting their systems to meet MISOE's needs, as described in the Journal.

The Consistent System

This report considers the adequacy of our software, known as The Consistent System, on the Multics time-sharing system, for use in MISOE.[1]

About Multics, the Computer

Multics is a large, advanced time-sharing system in use at the Massachusetts Institute of Technology. It was developed there jointly by M.I.T. and Honeywell (formerly GE) with some early design work also being done at Bell Labs. Multics has a number of special features which make it particularly suitable as a host to the MISOE application. These features include multiple access (over seventy simultaneous users can be accommodated); an exceptionally large file system which is always "on-line"; and an unparalleled ability to share files with other users, that is to permit (simultaneous) access to those designated, and to (dynamically) deny access to others. (As described below, this ironclad security is available for MISOE on a per-variable basis.)

---

[1] It is important to note that we comment only on the suitability of our software to perform for MISOE; we do not comment on the efficacy of MISOE in achieving its goals. Information about MISOE, on which the report has been based has been obtained from "Computer Assisted Planning for Education--a Prototype."(J. Res. & Devel. in Educ. 7, No. 2, 1974), and from an unpublished document entitled "Specifications for MISOE 'State 51' Feasibility Study".

The Consistent System -- the Software Library

It is upon the Multics system that we chose to assemble the Consistent System -- conceived of as the ultimate interactive computer facility for the behavioral scientist. The nature and extent of the data, and the way it must be treated serve to define behavioral science computing. First of all, there is a relatively large amount of primary data. Often the data derive from a survey where the initial collection cost is high -- that is, cost rises less than linearly with the number of questions asked, consequently many questions are asked. And, while costs do rise in proportion to the number of respondents, sampling variability dictates a respectably large population. Perhaps more significant than sheer size is the requirement that the identity of each datum be preserved.

Sometimes new attributes may be derived from combinations of those originally measured, and cases can be collapsed into sub-populations. Still, because of the multi-dimensionality of the data, and its unaccountable variability, the underlying elements must be preserved intact. Condensation with an acceptably low loss of information is not yet possible.

A second, important characteristic of behavioral science data has to do with the difference between the order of data collection and data analysis (the 'inverted file' problem). For instance, in an experiment where each subject serves as his own control, data is often analyzed by treatment, or condition; yet, it had been collected on a subject-by-subject basis. Likewise, any survey or multiple-dimensional testing situation collects a number of responses for each respondent on a per-respondent basis; these must often be analyzed on a per-variable, or per-question basis.

In concert, these factors dictate that much of behavioral science computing is 'data management'. Tools for data management and data analysis may be inextricable from one another. Because of this, behavioral scientists have so far been best served by relatively large and complicated 'packages' such as SPSS, Data text, Biomed, OSIRIS, and the like. These packages combine the capability for statistical analysis with a capability for data management. The reason for the success of these packages is the graceful intercorrections between the components. Outputs from one series of operations are properly constituted so as to serve as inputs to other operations.

Our objective was to provide, in an interactive environment, a loosely cohesive collection of programs all of which would work nicely together--that is, the cutput of any program should be already in the proper format to serve as input to any other program (at least wherever it made sense to the researcher). Now to the layman it naturally appears that this would certainly be the case. Of course, outputs from one program would be acceptable as inputs to another. Any other condition would surely be exceptional he thinks. However, incompatibility is more often the rule than the exception.

Because the CS was to be the repository of an extremely heterogeneous group of programs, little attempt was made initially to impose a standard syntax upon those programs. Thus the individual style of the programmer shows through to the user and already there is considerable diversity ...and an effort is underway to alleviate that diversity. The system already contains a large number of "programs", and the programs differ not only according to the style in which they expect their arguments, but also according to their size and function. Sometimes a single "program" is a single command

or logical request; sometimes a program is a collection of commands and opetions -- a small (or large) subsystem.

Yet another of our objectives was to produce a system that the average behavioral scientist could use without having to become a conventional programmer. Even without formal programming skills, the researcher was to be able to string together commands so as to merge them into a single, larger intellectual unit.

As you can see, the principles that guided our development make the Consistent System ideally suited to MISOE needs.

Data Management in the CS -- JANUS

One of the larger "programs" in the CS is JANUS. The JANUS system is an attempt to provide a complete, on-line data management environment to support data analysis and decision making. The system is designed to be directly usable by the administrator, researcher, statistician, or policy maker through the use of a problem-oriented language.

In addition to supporting the standard data-management functions such as data input, update, and display, JANUS provides a data transformation capability that enables the user to derive new data structures. This enables the user to answer questions from his data that cannot be answered with more conventional systems.

A particular view of data and data operations underlies the design of the JANUS system. The "data model" is the key to understanding JANUS.

A "dataset" may be conceived of as a row-by-column matrix (or as a card deck): the rows (or cards) are called "entities", and groups of columns are called "attributes". Suppose we had the following information:

#	Soc. Sec.	Last Name	First	School	IQ	reg	yr	wght
1	174-30-4567	Smith	Sam	votech	122	74	74	165
2	243-34-0293	Jones	Archie	ridge	*	73	74	227
3	187-56-8347	Weisheimer	Egbert	jepson	143	75	74	88
.								
.								
.								
10	174-30-5426	Kline	Harry	raditech	108	33	75	143
.								
.								
.								
72	236-63-0296	Hines	Bertie	ridge	*	76	73	127
.								
.								
.								
91	197-56-9137	Wishwasher	Dowbert	ridge	133	75	73	99
92	188-24-0292	Kahn	Harry	jepson	109	72	74	186
93	234-22-8347	Smarteimer	Herbert	jepsom	112	66	74	199
.								
.								
.								

We have a file of students (the entities) and a number of attributes for each entity: their social security number, their name, the school they attend(ed), their IQ, the region in which they now live, the year of their matriculation, and their weight. Note that the values which an attribute can take on depend upon the datatype of that attribute. Weight is a ration scale and is portrayed as an integer or floating point number. Year of matriculation is ordinal data and might be coded as integer data. The remaining variables are nominal data and are portrayed as numeric or text strings.

Each attribute can be stored as a separate Multics file. Consequently control as to which users can access the data is available on a per-variable basis. For example, name and social security number could have severely restricted access; the other information could be made generally available on a "read-only" basis, and a single individual might have "write" access in the event that certain information, found to be incorrect, had to be altered.

Note that missing values (denoted by *) are specially treated by JANUS. They need not be artificially coded as numeric values and are left out of all numeric computations.

JANUS allows retrieval of information conditional on the values of any combination of attributes. We can ask to

display all, for region = 73 & weight > 140

If we were only interested in the social security numbers and weights of those individuals we could ask to

display soc. sec. weight, for region = 73 & weight > 140

And, missing values may be used in conditional specifications, so we can ask to

display year, for weight = missing

And, of course, we can summarize any of the data by asking JANUS to

compute mean weight, for school = "jepsom"

Naturally, we could create new attributes in terms of old ones:

define-attribute stones as numeric = wght/14

Each variable so created could have its access controlled by the person who derived that variable.

Most important, JANUS allows one to create and use effectively new data sets related to the present one. Suppose an analysis was to be on a school-by-school basis. We would project this dataset into one where the entities were the schools. We would map into that derived dataset those attributes in which we were interested. Inasmuch as the mapping is a many:to:one mapping, we could aggregate or average the data. Thus we could end up with a dataset:

#	School	avge_wght	pupil_total
1	jepsom	177	66
2	raditech	12	55
3	ridge	153	91
4	votech	147	87

Moreover, we could infer data back from this dataset to our original and retrieve, for example, the names of individuals who attended schools with more than, say, 88 respondents in our sample (and where the average weight was greater than ... etc.).

If another dataset provided information about teachers (including the school at which they taught) then we could inter-relate the three datasets. We could, for example, retrieve the social security numbers of all students and teachers in the same school whose last names coincided (and where there more than so many respondents ... etc.).

In addition to its use for complicated, inter-related retrieval requests, JANUS can provide summary statistical information: Means, medians, standard deviations, distributions, scatterplots, regressions, and correlations. Furthermore, it can prepare data vectors for use by the wide variety of statistical and graphical analysis programs additionally available in the Consistent System.

Using the "Discourse" system a user could take data from JANUS and prepare a grid-map of the selected variables. Using a variety of distance metrics which Discourse can provide we could, for example, map the locations of pupils who chose not to attend schools closest to their homes. And, using supplementary data, we could superimpose geo-political boundaries, major arteries, etc.

For longitudinal data, TSP, the Time Series Processor, might be used. For cross-sectional data, separate analysis of variance routines in the Consistent System could be applied.

It should be emphasized that all of the capabilities indicated here are available now. They are available interactively at the terminal. Ample documentation is available on the system and its many components. Consulting help can be made available. All the user need supply is the data and an idea as to what kind of analysis he might like to apply. Only a fraction of the capabilities of the system have been reported on here and those but sketchily. Nonetheless, it is inevitable that some analysis tool or variant thereof, which a user might want, will be absent. Fortunately, the system was designed for growth by individual contributions; programming talent can be supplied if the user should need it (although experience has shown that he nearly never does.)

The general information given above should help the reader evaluate the MISOE problem and the Consistent System in relation to one another. There follows some brief answers to questions posed in a document entitled "Specifications for MISOE 'State 51' Feasibility Study".

The system can operate interactively in response to requests by the user to manipulate, retrieve and analyze specific data, as the above description of JANUS, brief as it is, serves to illustrate. The format of the commands are nearly exactly as shown (although abbreviations are also provided so as to lighten the typing load). Tabular reports can be generated with moderate control over the report format; additional editing capability is supplied separately in the system.

Graphic displays can be generated for the typewriter console, the line-printer, or Ards and Imlac CRT terminals. (In the event that such CRT terminals are to be used, we should have additional discussions; our tentative recommendation would be for one of the new, fast printing/plotting typewriter terminals, instead).

Off-line reports can be generated on the high-speed line printers at MIT, and, with suitable arrangement, can be mailed to an appropriate addressee. Input data can be in the form of "ASCII card images" (although with special arrangement, other formats would be acceptable). The media can be punched cards, or magnetic tape (formatted as ANSI-labeled tapes, industy st'd 9 track preferred.) Of course, data can be input from the terminal, and all data can be edited and altered at the terminal. Random number generation routines can supply realistic test data in the event that actual data is hard to come by for the demonstration phase.

As the JANUS discussion indicates, new variables can be generated and stored. They can be deleted, updated, or recreated. (Note, however, that in the case of related or derived datasets, updating of the derived datasets is not automatic as a result of updating the source datasets.)

Normal storage on the system is "on-line" and backed up on tape frequently (twice a week at least, and within an hour or two of any change to any file). This backup service is "free-of-charge" and the cost for such storage is $0.60 per 4Kbytes; storage can be made available without such automatic backup for about $0.20 per 4Kbytes and routines can be provided for manual backup. Retrieval from backup cost $10.00.

As to the weighting variables, one needs as an attribute the subpopulation relevant to that entity; from this one can derive new variables

as the product or quotient of original and count variables; the analysis would then be done on these weighted variables.

The command repertoire available to the user is reasonably simple, given the powers at the user's command. Simplification and tailoring of the language is easily accomplished if desired. A routine for "menu selection" is not currently available but could be provided if desired.

Within the JANUS data management system, the limit on a single dataset is about 100,000 entities (or rows) and 50,000 variables or attributes (columns). Because of the relations which can be established between datasets, the exact structure chosen will determine the "amount of data" which can be accommodated.

Using the Consistent System on MIT's Multics, there appears to be no reason that the MISOE "State 51" system should not be demonstrable in less than two months insofar if its specifications remain flexible. It is our opinion that JANUS software meets the spirit of every MISOE specification. It probably also meets the letter of every specification. However, a report such as this should not be construed as an ironclad guarantee.

The specifications given for the MISOE system are at once too detailed, and not detailed enough. This is not to criticize its authors; in fact, they really told their story well. However, we can never be sure of any point, whether it was meant literally or figuratively ... whether it was part of the narrative, or part of the specifications. To establish this would involve a lot of interaction with a number of MISOE designers. To carry out this end of the interaction, we would have to show how our system

handled the issue -- we would make up data and demonstrate ... and that, of course, is the "State 51" demonstration.

Then again, to make precise estimates of what a retrieval request might cost, we would need to specify exactly how it was qualified, from what population it was drawn, what relations were involved, etc. Again, to do this we would generate random data, interact with MISOE personnel so as to let them make the requests, and then look and see how much that request just cost. (Note, Multics provides accounting information on a per-command basis.) A table of rates for MIT's Multics machine is given at the end of this report. Again, we seem to be describing the "State 51" project.

Candidly, we know of no other system whose current capabilities so closely approximate the MISOE job. A guess as to cost for preparing and running the "State 51" demonstration indicates it could be done for less than $40,000 total cost. This is for five month's effort: two months of preparation, and three months of operation. We would supply manpower and computer time; MISOE personnel would participate and certify the effort continually.

We should emphasize that this is not a proposal. Moreover, MIT would be unlikely to accept a fixed price contract for the work. I have no doubt, however, that your agency could write an RFP (Request for Proposal) to which we could respond within those dollar-constraints. Moreover, I'm sure that satisfactory terms could be arrived at so as to (incrementally) assure your satisfaction. In our judgment, if you were to issue such an RFP, we know of no others who could respond credibly.

Rates for 6180 MULTICS

Interactive Usage

Shift	CPU	Memory	Connect	Terminal I/O
1	$300/hr	$18/Kmu	$1.25/hr	$0.00/K lines
2	$240/hr	$15/Kmu	$1.25/hr	$0.00/K lines
3	$120/hr	$ 8/Kmu	$1.25/hr	$0.00/K lines
4	$200/hr	$10/Kmu	$1.25/hr	$0.00/K lines

(The shifts run 9-6, 6-12, 12-8, and weekends, respectively)

Absentee Usage

Queue	CPU	Memory
1	$240/hr	$15/Kmu
2	$200/hr	$10/Kmu
3	$120/hr	$ 8/Kmu

I/O Daemon usage

Queue	Print	Punch*
1	$3.30/1000 lines	$5.50/1000 cards
2	$2.00/1000 lines	$4.00/1000 cards
3	$1.65/1000 lines	$3.30/1000 cards

Tape Usage

Tape Usage	$1.10/tape mounted + $25/hr
Tape slot	$1.00/month
Tape rental	$1.00/month

Disk File Usage

$0.02/page-day

Other charges

Connect time is $1.25/hr.
Connect time on 1200 baud channels will be $2.50/hr.

Retrievals not caused by system error will be charged $10.

*Currently punch is charged at same rate as print.

XSIM

The MISOE prototype system, "State 51", can be easily implemented by utilizing the existing features of the XSIM language on Interactive Data Corporation's computer facility.

The Basic System

The most important part of the prototype system would be a file system which is capable of efficiently storing and retrieving the various files in the Census Data System and the Sample Data System. Because of the volume of data involved, approximately 55 million bytes, the data would be stored in an ISAM-type file system which would be accessed directly by XSIM. Another important requirement for the prototype system is that data can be easily added, deleted, and updated. Moreover, it is preferrable that these data-editing operations be simple enough that they can be performed by people without prior computer knowledge.

Dynamics Associates can design and implement such a system in the XSIM language which will have all the features outlined in the MISOE specifications. The consulting cost for producing this system would range from $10,000 to $13,000. The computer cost would be approximately $2,000, and the cost of loading the initial data would be approximately $3,000. The cost of editing data subsequent to the initial loading would range from 5 cents to 75 cents per record.

The cost to store the system on-line would be approximately $5,000 per month. In order to avoid reloading costs, data stored off-line should remain on a disc pack rather than be transferred to tape. The cost of storage at IDC would be $20 per month. ($50 per month if the disc pack is supplied by IDC).

Analytical Capabilities

To insure the continued usefulness of the MISOE prototype system, it is very important that the system not only be able to recreate the analyses shown in the MISOE specifications but also be flexible and powerful enough to allow further analyses to be easily performed on the data. A System programmed in XSIM would have access to all the statistical routines (including regression) and the graphing and report-generating capabilities of the language, thus providing an ideal environment for analytical studies.

To reproduce the tables in the Journal of Research, it is recommended that intermediate data records be used to minimize calculations which could be shared by similar analyses. The cost of calculating the intermediate records for a set of analyses would range from $25 to $100, and those intermediate records would only have to be recalculated when the data files were updated. Given the existence of the intermediate records, the cost of producing a report similar to Table 10 would be approximately $1, while a report similar to Table 22 would be $5. (All other tables would fall within this range.) Dynamics can write the programs necessary to calculate the intermediate records and produce the various tables in Chapter 8 of the Journal for a cost ranging from $500 to $1,250 per table (inclusive of computer time). However, at the option of MISOE, Dynamics would be willing to train, free of charge, MISOE personnel to program additional analyses on their own.

Dynamics feels very strongly that the MISOE prototype system will be greatly aided in meeting its stated objectives if it is implemented in the XSIM language by Dynamics.

Response to Proposal (I)

Professor John J. Donovan, M.I.T.
Associate Professor of Management Science

This report is to asses the strategy, requirements, and alternatives for a computer information system for MISOE.

Let me first state two inherent problems of such an information system that influenced my thinking and recommendations. They are speed of implementation of a prototype and flexibility. That is, for further funding, visibility, and continued success, I believe it is imperative that a prototype using realistic data be running in a matter of months.

Secondly, I believe that while the system requirements are specified far better than any other system I have seen at its stage of development, the system requirements will change. These changes will occur as the problems facing the states change, as the data sources and types change, and as the perception and capabilities of the systems evolve.

A list of general facilities that an information system must have to meet MISOE needs are:

1) Speed of implementation
2) Flexibility to change the types of queries
3) Flexibility to change the structure of data
4) Flexibility to change the syntax of the query language
5) Flexible modeling facilities
6) Protection and security mechanisms
7) Data validation facilities
8) Flexible capabilities for storing and manipulating different types of data

As to the assessment of both XSIM and JANUS, let me say that either choice would be satisfactory. On technical grounds JANUS is perhaps more suited, as it was conceived, designed and implemented for applications such as MISOE. Thus, I would give my technical recommendation to JANUS. As this report will discuss, however, XSIM may have some non-technical characteristics that may make it more suitable, such as being on IBM equipment and having a staff that is really anxious to acquire the business.

JANUS was designed at M.I.T. to be used for storing, retrieving and analyzing statistical data. XSIM is a spinoff from the TROLL modeling facility developed at M.I.T. in the National Bureau of Economic Research and was initially designed primarily as a facility for economic modeling. Hence, emphasis was placed on facilities for time series, regression analysis, etc.

The major difference between the two systems lies in the different emphasis of their initial purpose. TROLL has good modeling facilities but poor data management facilities (e.g., no general way to store and manipulate non-numeric data, difficult to restructure files, etc.). JANUS has good data management facilities but poor modeling facilities.

Both systems are making additions to make up for their respective differences. For example, a modeling facility (TSP) has been added to JANUS.

Both systems have the following disadvantages:

- No ability to monitor access to files and restructure files for efficient organization
- No general facilities in the system for protection. (Note: both systems are imbedded in operating systems that have protection facilities). Facilities such as audit traces of user's reading data are not available in TROLL or JANUS.

More specifically, the advantages and disadvantages are as follows:

JANUS

Advantages	Disadvantages
a) Used by the State of Massachusetts Department of Manpower	a) Is in an academic environment
b) Designed for such applications	b) It is on Honeywell equipment
c) Capable of building the system very quickly	
d) Operating system has more security features	
e) Has modeling facilities	
f) Has good graphic capabilities	

XSIM

a) On IBM equipment

b) DYNAMICS is eager to acquire business

c) Good modeling facilities

d) Graphic capability

Response to Proposal (2)

Maurice Shirley, MISOE Staff

The Cambridge Project and Dynamics Associate, Inc. returned somewhat similar reports indicating their systems could meet the requirements of MISOE "State 51" and at a resonable cost but not specifying such costs in minute detail.

The major difficulty in arriving at detailed cost estimates for the MISOE prototype system is the extremely flexible nature of the actual operation of the system. We want the user to have the option of analyzing any of the items in our large data files in an almost infinite number of ways.

The only way we could predict the exact cost of operating "State 51" would be to specify exactly what the prospective users at workshops will be allowed to do while interacting with the data files. This is not what we want to do so our cost estimates in this area rely heavily on the judgment and experience of the developers of the proposed software.

The JANUS data management system stores data in matrix type files called data sets in which each row is an individual record and the columns are variables in these records. In contrast, the XSIM system stores its data in time-series in which each period in the series represents a variable. A file in XSIM can be considered as a group of series.

In both systems statistical routines can be performed without reformatting the data.

In JANUS the names of variables are limited to 6 characters the first of which must be alphabetic. XSIM limits the names of variables to 8 characters the first one being alphabetic. The MISOE DATA INDEX would be the same for either system.

To illustrate the user command languages of JANUS and XSIM, consider Table 5, page 84 of the JOURNAL. This Table is simply a comparison of the average starting incomes of OE program completors who did participate in cooperative study programs versus the average starting incomes of OE program completors who did not participate in cooperative study programs.

For JANUS the data would be stored in a DATASET (named IMPACT in this case) where each row (or record) is called an entity and each column (or variable) is called an attribute.

ENTITY	PID	START	COMPLT	TYPE	COOP
1	001	6000.	1	1	1
2	002	5600.	0	0	0
3	003	7000.	1	1	0

In this example the column PID contains the student's identification number, START contains the student's starting salary, COMPLT is Coded 1 if the student completed the program and 0 if not, TYPE is coded 1 for OE programs and 0 for NOE programs, and COOP is coded 1 for participation in a cooperative study program and 0 for no participation.

The commands to find the required means would be:

COMPUTE MEAN OF START, FOR COMPLT = 1 & TYPE = 1 & COOP = 1

COMPUTE MEAN OF START, FOR COMPLT = 1 & TYPE = 1 & COOP = 0

For XSIM the data would be stored in a GROUP (named IMPACT in this case) where each row (or record) is called a series and each column (or variable) is called a period of that series. The column titles and data are the same.

SERIES	PID	START	COMPLT	TYPE	COOP
IM 1	001	6000.	1	1	1
IM 2	002	5600.	0	0	0
IM 3	003	7000.	1	1	0

The commands to find the required means would be:

```
DO MSK1 = IF COMPLT EQ 1 AND TYPE EQ 1 AND COOP EQ 1 THEN 1 ELSE 0;
DO GRPBYMSK. (MSK1,$'IMPACT,$'IMPACT,$'IMPAK1);
DO SAVE      MEAN.('IMPAK1,2);
DO MSK2 = IF COMPLT EQ 1 AND TYPE EQ 1 AND COOP EQ 0 THEN 1 ELSE 0;
DO GRPBYMSK.(MSK2,$'IMPACT,$'IMPAK2);
DO SAVE      MEAN.('IMPAK2,2);
```

The matrix type files and the syntax of the instructions in the Consistent System (JANUS) seem especially suited to the MISOE system with its emphasis on cross-sectional analysis.

Both XSIM and JANUS can store the above series of commands in a macro file and repeat the series by issuing only one command at the terminal.

The Consistent System (JANUS) now has ARDS graphics software in operation requiring either an ARDS terminal or an IMLAC minicomputer which can simulate the ARDS terminal. Other graphics software such as Tektronix could be added to the system if desired. For their own use the developers of the Consistent System are at this time installing several SYSTEM 75 terminals which use the Diablo Hytype I printer and have graphic capabilities.

The XSIM software can at present do bar charts and scatter plots but not line graphs. Here again, graphics software such as Tektronix could be added to the system. For nongraphic output either system will accomodate a wide range of terminals including low cost typewriter terminals, as well as CRT terminals.

Both systems have high-speed line printers available for off-line reports. The XSIM system seems to be somewhat more flexible than JANUS in formatting reports for the terminal (or the off-line printer) but the JANUS developers assure us they will improve their report generating routines if we find them inadequate for the prototype system.

In both systems, data can be typed in at the terminal or loaded from punched cards or from tapes. The data has to be properly formatted for the particular system but this should not be a problem. The prototype system will probably have a small amount of data keypunched or typed in at a terminal and then modified and expanded to complete the data files. Both systems should handle this with no difficulty.

In regard to dynamic simulation, DYNAMO is available along with XSIM on Interactive Data Corporation's system. We would have to prepare a special data file from the XSIM files and then read it into the DYNAMO system rather than access the prototype files directly. DYNAMO is also available on the MULTICS system at M.I.T. Here again we would have to prepare a special data file for the DYNAMO system.

At the present time, the JANUS system is available only on a Honeywell 6180 computer. It is expected that in the near future a version of JANUS for the IBM-370 computer will become available. Until such time, we would be restricted to using M.I.T.'s computer since there are a limited number of 6180's in existence and the others are unlikely to be available to MISOE. XSIM, on the other hand, is now in use on IDC's IBM 360/370 and can also be purchased outright for transfer to other IBM 360/370 computers.

The method of weighting the SDS data is not clearly specified in the Journal other than to say that all statistical analysis shall be on weighted data. If the intent is to convert a file of sample data to a file of weighted data before performing a statistical routine, then both XSIM and JANUS can create such intermediate files by means of user-created routines. While such weighting routines could be easily stored in the system for later

use, as a practical matter, we would need a different weighting routine for each change in criteria in selecting data for analysis. If the intent in using weighted data is to have the statistical routines automatically weight the sample data during analysis, then it appears that the statistical routines in either XSIM or JANUS will have to be rewritten. However, alternatives might be "discovered" as we move into "State 51". MISOE would have to retain the "unweighted" routines for analysis of CDS data.

XSIM suggests MISOE purchase or rent a mountable disk pack (IBM 3336) which can be stored off-line for a nominal fee and mounted by the operator on short notice. The JANUS data is backed up by frequent tape "dumps". These tapes are stored off-line for free with a nominal fee for reloading to the on-line disk packs. However, it would require 24 hours' notice to reload our data files. It is assumed that the MISOE requirements are to have the data continuously on-line during the several weeks of demonstrations and then stored off-line, easily available for later demonstrations.

Cost Estimates for Software

Table 2 is a summary of the expected computer costs for the MISOE prototype system using either JANUS (the Consistent System) or XSIM. It does not include salaries for the MISOE data processing staff or computer terminal costs which are discussed separately.

These costs are for one complete SDS cohort as described in the JOURNAL (15,000 students with a complete set of IPPI batteries) but with a truncated version of CDS (described earlier in this report). The estimates are based on the reports submitted by the Cambridge Project and Dynamics Associates, Inc. and on several follow-up discussions with their representatives.

In particular:

1. Initial Design and Set-up - Includes both manpower and computer time to be supplied by the developers to set up the initial system, load the data index, and do some programming to custom tailor their systems to MISOE's particular needs. The MISOE staff would be working closely with the developers at this stage and would assume part of the work load. The estimated cost for this is approximately the same for both JANUS and XSIM.

2. Keypunch Data - The assumption here is that a small amount of "real" data (possibly for 100 students, for example) will be prepared by the MISOE staff from test data already on hand and loaded into the system from punched cards. This data would then be expended by altering certain key items in each record before copying the altered record to enlarge the data files. In this way, we can generate any size data base we desire (or can afford) with "realistic" data while keeping keypunching costs to a minimum. The exact details of how "real" the "State 51" data has to be have yet to be determined. Until this is done we cannot attempt to estimate the time and cost of specifying and describing the "State 51" data, but acknowledge that the data must be completely specified before we can begin implementing the interactive system. In the event the assumed method of data generation is considered unsatisfactory there will probably be an increase in the time and expense involved in implementing the system.

3. Loading and Copying Data - This represents the computer costs to load the original (keypunched) data, alter and copy the records until we have the complete "State 51" data base. It also includes storage costs during this process. The JANUS cost estimate is lower here because of lower CPU charges.

4. On-line Storage - This assumes that the entire "State 51" data base will be stored on-line for 12 weeks. This item could possibly be reduced if there are several consecutive days during this period when we were not using the system so that we could go to off-line storage. XSIM has a slight cost edge over JANUS here.

5. Operating Costs - This was the most difficult cost to estimate because the parameters of the demonstrations and workshops have yet to be specified (and probably never will be because of the desired flexibility" of the interactive process between users and data). The assumption here was that over 12 weeks approximately 200 reports similar to those in Chapter 8 of the JOURNAL would be generated. Here JANUS has a cost edge because of lower CPU charges.

6. Telephone Service - The cost of telephone service cannot be ignored. Assume the demonstrations are held in Lowell and we employ several terminals (at one telephone line for each terminal).

7. Miscellaneous - There will always be some editing, debugging and last minute changes required. The cost is difficult to estimate but we should allow some leeway here for unexpected problems.

The total cost estimates in Table 2 are approximately 50 per

cent over our tentative budget of $40,000 for computer costs. Bear in mind, however, that these figures are for a complete SDS and for a CDS, that, while truncated, still includes an enormous array of data. Reducing the size of "State 51" to one half the size of Massachusetts would also reduce the cost of items 3, 4, and 5 in Table 2 and would reduce our total computer costs to approximately $45,000. Truncating some of the SDS batteries (particularly the INPUT battery) would also reduce the costs of items 3, 4 and 5 in Table 2 and possibly item 1 as well, to achieve a total computer cost of about $40,000. These proposed reductions would still provide a large enough data base for the purpose of the learning laboratory phase of Collaborative MISOE.

The costs listed in Table 2 are reasonable estimates but it should be pointed out that a major benefit of implementing the prototype system will be obtaining actual operating costs which can be projected to real MISOE.

TABLE 2
Prototype System Cost Estimate

	JANUS	XSIM
1. Initial Design and Setup	$16,000	$15,000
2. Keypunch Data	1,000	1,000
3. Load and Copy Data	13,000	19,000
4. On-line Storage	12,000	10,000
5. Operating Costs	12,000	15,000
6. Telephone Service	1,000	1,000
7. Miscellaneous	4,000	4,000
TOTAL	$59,000	$65,000

Terminal Configuration Costs

The MISOE interactive system is designed as a remote time-sharing system with the user operating a terminal connected to the computer by a telephone line. There are a multitude of computer terminals on the market today and most of them could conceivably be used by MISOE. A typical terminal allows the user to transmit messages to the computer from a keyboard and either prints the communication betwen the user and computer on paper or displays it on a CRT (cathode ray tube). The choice between a CRT display and a printer display is usually determined by whether the user needs a printed copy of the data or is only interested in looking at or scanning the data.

CRT terminals can usually operate at faster transmission speeds and are quieter than printer terminals. They are also considerably more expensive and can display only a limited amount of data at a given time. A device for making a paper copy of the CRT display can be added at extra cost. If this copy is not made the data or messages displayed will eventually be lost to the user. Printer terminals provide a complete record of all communications between the user and the computer and can handle long reports that would be too large for the screen of a CRT terminal.

For the MISOE Prototype System which will be demonstrated to groups of people, a CRT terminal with as large a "TV" screen as possible is very desirable. For a real MISOE interactive system, a combination of CRT and lower cost printer terminals could prove more efficient, since the many users might want a printed copy of all communications with the system.

In Table 3, we list the purchase and rental prices of several terminal configurations which we believe are appropriate for the MISOE Prototype System, and in the following paragraphs describe some of their features.

The SYSTEM 75 terminal is a typewriter terminal using the Diablo Hytype I printer and having graphics capabilities. It uses standard width computer paper and can print 132 (or 158) characters per line. It can send and receive at 30 characters per second which is three times as fast as the Teletype 33 which we now have.

The TECHTRAN cassette unit allows the user to prepare, edit and store data or programs off-line for later transmission to the computer. The TECHTRAN 4100 and the SYSTEM 75 together make a complete terminal system able to operate off-line as well as send and receive to and from the computer. This cassette unit can be used with other terminals, including CRT terminals.

The Tektronix 4010-1 and 4015 are CRT terminals having graphics capabilities. At this time, there is a 6 months waiting period for delivery of the large (19") screen 4015. The Tektronix 4010-1 and 611 monitor both have smaller 6" by 9" screens but delivery should not be a problem. A 4630 hard-copy unit can be used with either Tektronix terminal. We might also use the Tektronix graphics software if we choose their terminals (particularly if we use XSIM which does not have line graphs at present).

The IMLAC PDS-1D is a minicomputer with a CRT display. Many optional features are available and it has graphics and data editing capabilities, however, it is a high cost configuration.

The ADDS CONSUL 980 is a CRT terminal which the MISOE staff has not seen. It features a black on white display which we hesitate to recommend without a demonstration. The other CRT's listed in Table 3 all display light-colored letters on a dark background. The ADDS 980 does have good data editing capabilities and also a graphics capability which we believe to be

inferior to the Tektronix graphics. It is not recommended at this time, but after a demonstration it could be our final choice because of its low cost and the possibility of using a TV set monitor with it.

The Teletype Model 38 is a wide carriage (it uses standard computer paper) verios of the Model 33 MISOE now has in stock.

The TRENDATA Model 2000 uses an IBM Selectric typewriter to provide a low cost printer terminal.

Acoustic couplers are a necessary interface between the terminal and a telephone set for all terminals described above.

TABLE 3

Terminal Configuration Costs

HARDWARE	PRICE	ANNUAL RENTAL
SYSTEM 75 (Printer Terminal)	$5500	$2300
TECHTRAN 4100 (Cassette Unit)	$2500	$1100
TEKTRONIX 4010-1 (CRT Terminal) with 611 CRT Monitor	$8500	$5000
TEKTRONIX 4015 (CRT Terminal)	$9500	$6600
IMLAC PDS-1D (Minicomputer with CRT Terminal, 21" CRT Monitor and other accessories)	$26,000	$10,400
ADDS CONSUL 980 (CRT Terminal)	$3000	$1500
TEKTRONIX 4630 (Hard Copy Unit)	$5000	$2500
Teletype Model 38 with Wide Carriage (TTY Terminal)	$2250	$1000
TRENDATA Model 2000 (Printer Terminal)	----	$1200
Acoustic Coupler	$500	$180

NOTE: Prices include installation charges.

Recommendations

The cost estimates in this report indicate that "State 51" can be implemented within our proposed budget if we limit the size of the data base. There is really no reason not to do this since assembly of a truly enormous data base designed primarily for learning purposes is not an efficient use of funds. The reduction in data base size would still leave a data base large enough to examine the flexibility of MISOE and to specify the operational costs of a real MISOE. These cost estimates indicate that the two proposed software systems, JANUS and XSIM, are in approximately the same price range, with the JANUS system expected to have only slightly lower costs for implementing "State 51".

While either JANUS or XSIM would be adequate for the MISOE interactive system we recommend using JANUS for the following reasons:

1) The system was designed for applications such as the Management Information System for Occupational Education (MISOE). It is closer to being the "ideal" system described in the JOURNAL than any of the other data management systems we have investigated.

2) It has data management capabilities which are more suitable for the MISOE system with its large amounts of cross-sectional data. While MISOE places great emphasis in having available a sophisticated statistical analysis capability, the primary requirement of its interactive system is the ability to easily and quickly "manipulate" data, that is to select, screen, sort, rank, count, sum, etc., variables and to create new variables from existing files. The JANUS file structure, and more importantly, its user commands are simpler and easier to use.

3) There is a predicted cost advantage for using JANUS although not a large one.

Professor Donovan has noted two disadvantages of using the JANUS software. The first, that JANUS is now available only in an academic environment rather than being supported by a profit-oriented business concern, is difficult to assess. The MISOE staff has already contacted several organizations who are now using the JANUS software and we plan to meet with them shortly. Their experiences in this area should be a reliable guide to possible difficulties.

The second disadvantage, that JANUS is only operational on a Honeywell 6180 computer, should present no difficulty during the life of "State 51", but could be a problem for "real" MISOE which is expected to encompass several states with correspondingly large data bases). In the event that JANUS does not become transferrable to an IBM computer (or others) in about a year, there would be no insurmountable difficulties in changing to XSIM or to any other software suitable for MISOE which might then be available.

In regard to terminals for "State 51", the recommendation is that MISOE acquire a SYSTEM 75 terminal with a TECHTRAN 4100 Cassette Unit as soon as we are ready to start implementing the system. This equipment would be used to set up, test and debug the system and also to store and edit data off-line. As our primary terminal, it would also be used for processing the Massachusetts "Fall Report" and IMPACT study.

For MISOE workshops we should acquire a Tektronix 4010-1 with a 611 CRT monitor. We also need at this time two teletype Model 38 terminals for use in the workshops. This would enable us to divide our participants into four groups for the workshops (2 groups using terminals without graphics).

Leasing of terminals is recommended rather than purchase for the following reasons:

a) Leasing agency will be responsible for maintenance.

b) We can update our equipment periodically according to our needs which are expected to change fairly often or as improved equipment becomes available.

c) In the event our initial choice of equipment is not satisfactory, we will not be "stuck" with unusable "lemons".

d) Changing computers at some future date might necessitate a change in terminals.

The recommended configuration of one SYSTEM 75 with cassette unit, one Tektronix 4010-1 with Monitor and two TTY 38 units plus 4 acoustic couplers and other accessories can be leased on an annual basis for $13,000 per year.

To sum up, we recommend using the JANUS software on M.I.T.'s MULTICS system with the terminal configuration described in the preceding paragraph. The "State 51" data files should represent a state approximately one half the size of Massachusetts with a truncated version of the SDS batteries. Our investigations lead us to believe that "State 51" can be implemented and demonstrated for eight weeks at a cost of $40,000 for computer expenses plus $13,000 for terminals. We expect the prototype could be ready for demonstrations within two months after the initial data became available. This assumes that the "dummy" data is created as described earlier in this report and that we use the JANUS software as it now exists or with very little modification. The costs of processing the Massachusetts "Fall Report" and IMPACT study on this system are not included in the "State 51" estimates.

APPENDIX I

Analysis Capabilities

of the

Consistent and XSIM Systems for MISOE

Dr. John A. Creager

The MISOE staff has examined certain available software packages to determine their capability for handling such MISOE requirements as creating and maintaining an extensive data base, and for their ability to perform and report results of specified analyses of data in a user interactive environment. Although many software packages will do parts of the MISOE task, it is difficult to find software directly usable or readily adaptable to the whole project. To the extent that a single package meets most of the project requirements and can be readily modified or extended to meet the rest, such a package is preferable to interfacing several smaller packages developed by separate sponsors. MISOE staff narrowed the search to three candidates and sought proposals from the software sponsors for possible application in the "State 51" MISOE Prototype system. One of the sponsors declined on the grounds that their package would not be adequate to meet MISOE requirements or be adaptable at reasonable cost. This report comments on the ability of the two remaining packages, The Consistent System (CS) and XSIM, to meet the statistical, i.e., the data analysis, needs of MISOE. It is assumed that if MISOE needs can be reasonably met, those for the prototype system can be. Hence, in commenting on statistical adequacy of software, my remarks are focused on fully operational MISOE.

Excluded from my comments are remarks about interfacing hardware and software systems or any possible interactions between such problems and the analytical adequacy of the software. The materials provided by the sponsors have quite typical limitations for assessing (1) the degree of numerical accuracy in application of particular routines, (2) the efficiency in computer time and costs, and (3) capacity in terms of the number of entities and the number of variables each routine can handle. The ensuing comments are essentially concerned with the presence or absence of analytic capabilities required by MISOE, and are made under five headings:

1. Consistent and XSIM vs. MISOE's Operations and Output Indexes
2. The Missing Data Problem
3. Weighted Analysis
4. Crosstabulation
5. Regression: Multiple, Stepwise and Otherwise.

Consistent and XSIM vs. MISOE's Operations and Output Indexes

In checking the software package documentation as available against Table 3 of Chapter 7, of the MISOE specifications described in the Journal of Research and Development in Education, Vol. 7, No. 2, 1974, I find both packages to have the basic logical and arithmetic operators available for use in generating variables and as parameters of macro-commands. The ability of both packages to display frequency distributions presumes basic counting and summing operations. The notion of creating and storing intermediate data records containing sums, sums of squares and cross-products to save costs of their recomputation for repeated calls with the same summation indexes was indicated in the XSIM proposal. This point should be explored during

prototype studies to ascertain the gain against the fact that there are many different sectors of aggregation possible in MISOE.

I did not see any explicit reference to sorting, ranking or centile scaling in CS documents. Sorting and a rank function are explicitly mentioned in XSIM. Presumably ranking and scaling operations are possible in CS through a computation and display of frequency distributions and the creation of derived data sets.

Basic distributional statistics are available in both packages, except for the mode, which MISOE may want in dealing with missing data imputation for categorical variables. Again, this appears to be possible by initial display of the frequency distribution, from which the modal category can be identified and its code be ascertained. Measures of relationship will be discussed in ensuing sections.

Although both systems appear to provide options for output of results at either a terminal or a line printer, it is far from clear that CS could generate displays of the form and content shown in Chapter 8 of the Journal article, where, e.g., the rows are defined by different variables rather than by different levels of the same variable. XSIM shows examples of such displays from its report generator. It is also doubtful that the neat display of targets and criteria would come out in the table headings or that variable and group names could be adequately displayed by CS. Some of these matters are merely aesthetic; assuming the analyses can be performed, the results may be displayed in quite useable forms without conforming to the Journal examples in all details.

Generally, both packages seem to provide the basic analysis capability required. The XSIM proposal claims DYNAMO adaptability; it is not

clear whether or how, CS might interface with DYNAMO. Nevertheless, CS appears to be a powerful and flexible package for database management, analysis, and retrieval, especially if its sponsors are willing and able to fulfill expectations for MISOE not currently available. Both packages appear able to handle some of the more sophisticated aspects of econometric analysis and time series data. The time series data should be viewed as a special subcase of longitudinal data in which the same variable (or set of variables) is subjected to repeated measurements over time, thus generating highly multicollinear data sets. MISOE plans to have some of this kind of measurement in its total longitudinal data base. The TSP of CS, although apparently designed with emphasis on time series data processing, appears to be able to handle many aspects of the more general longitudinal processing requirements.

The Missing Data Problem

I found no reference to the treatment of missing data in XSIM documentation. JANUS, however, proposes to code an asterisk and then ignore such data in all numeric computations. In regression, cross products are computed ignoring entities with either value missing. I believe this to be unsatisfactory for MISOE, preferring the use of imputed values. I gather, however, that JANUS could handle the matter by using the asterisks in the initial raw data sets, to be used by level 2 and level 3 users, performing the analyses necessary to identify the imputed values and the records to which they must be imputed in derived data sets used in analysis. I assume it would be a minor problem to go from the tapes resulting from optical scanning in which missing data would probably be blank (unless the scanner were otherwise programmed) to initial JANUS datasets.

Weighted Analysis

Since most analysis in SDS is to have reference to MISOE populations and subpopulations, MISOE development and planning provided for the derivation and addition of sampling weights to basic data files as they are created. The expectation was that software would be available that would automatically weight the data during summing and counting operations in the various statistical routines. Except for the apparent availability of weighted regression in the TSP, noted in a later section, this does not appear to be the case with either CS or XSIM packages. With both packages, however, there appears to be the capability of obtaining desired weighted results through redefined (recoded) attributes using the appropriate arithmetic functions. Properly done, this would be followed by calling the appropriate statistical routine.

From the information provided, it is far from clear just what would be involved in actual practice over total MISOE with its multiplicity of subpopulations, files and weights. My present judgement in terms of cost and efficiency is that, in the long run, it would be more appropriate to make selective adjustments in the most critical routines, such as those producing basic distributions and distribution statistics. This could be done in such a way that weights could be preset to 1.0 rather than read from datasets in applications to CDS information.

In order to determine more precisely what is involved and the tradeoffs of adjusting our thinking and operations to fit the packages versus adaptation of routines, the sponsors of the packages should be queried as to how they would handle a set of specific examples for generating and

displaying weighted distributions, crosstabulations, univariate statistics, and contingency measures (e.g., phi and chi-sq.).[1]

It should be noted that we need to "normalize" all weights such that their sum over any particular cell equals the population N of that cell; otherwise, a different weighted mean may result if it is computed as sum WX/N rather than as sum WX/ sum W, within cells and/or aggregated across cells. This also applies to derived weights which are products of particular weights, and is more critical if the software divides by N rather than sum W.

Given the mathematical and other technical sophistication demonstrated in these software package documents, I am convinced they can be made to handle MISOE needs for weighted analysis; the question is one of overall cost and convenience to MISOE and its various levels of users.

Crosstabulation

Both statistical packages appear able to generate and display crosstabulated data. The XSIM documents give little indication of just how this is controlled by the user. I was unable to find any explicit reference to crosstab statistics such as phi or chi-square in XSIM.

JANUS has an explicit command for computing crosstabs with chi-square and phi-square computed and displayed. There is no indication that anything other than these statistics and the cell frequencies are displayed. Most users like to see displayed with cell frequencies, the row percentages, column percentages, and/or percentages against the grand total. It should also be noted that this command apparently assumes unweighted or pre-weighted frequencies. Since the attributes to be tabulated must be in nominal or

[1] It should be noted that maximum and minimum values are not affected by weighting and are necessarily those of the sample even if displayed with a weighted mean.

serial integer codes, an attribute like income or IQ must be recoded into range groups by the user before calling for a crosstab; this will not be likely to bother level 2 and level 3 users, but is the kind of thing that raises questions about the system sophistication required by the level 1 user.

Without seeing examples of displays of crosstabs, it is difficult to tell how well the numerical levels of attributes are identified in CS: just by numeric codes or also labeled; are attribute names also displayed? I raise as different questions: (1) whether a software system can generate meaningful, desired analysis; (2) the convenience to the user in making the right input calls: and (3) the quality of output displays.

JANUS does have an important feature for crosstabs: the ability to specify restricting conditions. This feature can be used to generate higher dimensional tabulations, e.g., A x B x sex could be generated by first calling for A x B for males and then for females; if desired, an additional tabulation without the restriction can be called. Again, however, if one wanted cell percentages against any of the marginal totals or marginal percentages against the grand total, additional ad hoc cells and operations appear to be required.

Regression: Multiple, Stepwise and Otherwise

Both software packages have some kinds of multiple regression capability. Both seem to enter regression with previously developed correlation matrices. This is probably most desirable for MISOE provided that R-matrices computed over major subpopulations are saved for reuse and can be directly called as regression inputs by the variety of dispersed users. Both packages appear to have the capability of generating such matrices as needed,

i.e., ad hoc where a saved matrix is not adequate. The regression packages permit selection of a subset of variables from the R-matrix for particular runs. What is not clear from the package documents is the capacity of these correlation matrix and regression routines in terms of sample sizes and the number of variables that can be included. This is important in MISOE with its large datasets and possibly large group sizes. The examples shown in the documentation of both packages have small sample sizes and numbers of variables. This matter needs further clarification.

Both packages have the simple multiple regression capability, wherein a fixed set of independent variables is forced into the analysis. The simple, fixed multiple regression is appropriate where the user already has a good grasp of the data base (or it is very small) and clear hypotheses. I do not see a capacity for weighted correlation matrices or regression in XSIM, although the TSP has such capabilities in CS.

In planning the analysis system for MISOE, emphasis was placed on having a generalized stepwise regression capability, which permitted selection of highly efficient subsets of variables to enter regression either freely, or in combinations of fixed and freely entered subsets. Moreover, it is often desirable in exploring a large data system to have a display of the situation at each step of a stepwise regression analysis so that one may judge whether capitalizing on sampling errors may be distorting results. XSIM has a stepwise procedure (unweighted) but does not appear to display results at each step without making a series of calls, each one conditional on the results of prior calls. Although the TSP has a number of regression options (weighted, Bayesian, nonlinear, "two-stage", and crossproduct-based) which provide much flexibility, it has no explicit stepwise option. Again,

it may be possible to accomplish the desired analyses given user sophistication with the software and by multiple, serialized calls involving conceptual adaptation of the two-stage routine. Both packages appear to be able to compute, store, and operate on residual scores, and to plot regressions.

In displaying regression results, both CS and XSIM will display R^2, metric coefficients, the constant, and significance information. Users wanting standard coefficients for interpretation and other operations involved in some kinds of "relative contribution" analysis would have to collect the metric coefficients and standard deviations in vectors and convert to a vector of standard coefficients. The metric coefficients are most relevant for computing predicted values and residuals.

Epilogue

Despite some of the matters noted above, I am generally impressed with the generality and flexibility of these packages. Overall, I am inclined to favor CS over XSIM, based on the documents in hand, and a presumed willingness of the sponsors to facilitate and expedite adjustments where necessary to meet MISOE requirements. XSIM, however, appears to have superior output display and graphics capability and explicit stepwise regression capability. CS appears to have the broader range of analysis capabilities and tremendous flexibility, which will be further enhanced if certain matters claimed to be under development are satisfactorily completed in reasonable time. I retain some misgivings regarding the degree of sophistication and patience required of the general user. What further user-oriented documentation with appropriate examples worked through and results

displayed can be obtained or developed? I have the feeling that considerable insight and practice are required before the typical user would feel comfortable at the terminal.

Some of my misgivings and criticisms may be clarified by further discussion with software sponsors before final MISOE commitment to application in the "State 51" Prototype Study. Others may be resolved by further development with a chosen package as part of the Prototype Study itself. Operational MISOE must have a completed and fully satisfactory software system on a "ready-to-go" basis with full and illustrative documentation.

APPENDIX II

Cost Projections to Real MISOE

Cost estimates thus far in this report have been for an interactive system loaded with "dummy" data for laboratory "State 51". Table 4 contains estimates of the data processing costs for a real world MISOE system for Massachusetts. These estimates include the cost of temporary clerical workers where required but does not include the salaries of the MISOE data processing staff. The costs of printing, distributing and collecting of forms are not included. The following brief descriptions of the cost items in Table 4 assume that data processing costs begin when the completed CDS and SDS forms are returned to MISOE.

A. Decollate SDS Data Forms. The returned forms must be manually sorted, separated and checked before Opscanning.

B. SDS Opscanning. This assumes that all SDS forms including IMPACT are OPSCAN.

C. Edit SDS Tapes. We plan to edit the tapes produced by the scanner on a small computer before loading to the interactive system.

D. Set up and load SDS files. Set up files in the interactive system and load SDS data.

E. Decollate CDS Data Forms. The returned forms must be checked and sorted manually before keypunching.

F. Keypunch CDS Data. For each year of CDS, approximately 20 million characters of data must be keypunched. This presumably replaces some of the Department of Education's present keypunch workload but much of it is additional work.

G. Set up and load CDS Files. Set up files on the interactive system and load CDS data.

H. Check and Edit CDS Data. The amount of work here will depend on how accurate the keypunchers were. If all keypunch work is verified this task could be minimal.

I. Operate System. These are computer costs to analyze the data and generate reports and includes storage charges while the data is on-line. Any estimate of how much and how often OE managers will use a real live interactive MISOE system has to be a "wild guess" at this time. There are certain reports (such as the Federal Report) that must be prepared from the CDS files each year. Other reports and analysis of the SDS data would be at the manager's option. It would obviously be excessively expensive to have the entire MISOE data base on-line ready for instant use 365 days a year. A reasonable estimate would be about two months "on-line policy making" per year. These on-line periods would be scattered throughout the year, a few days or a few weeks at a time, as needed. Our cost estimate is for 10 weeks on-line for both CDS and SDS.

J. Terminal Rental. Here our estimate allows for 6 terminals of the SYSTEM 75 type with 2 of them having cassette units.

TABLE 4

ESTIMATED ANNUAL DATA PROCESSING COSTS

FOR REAL MISOE FOR MASSACHUSETTS

		COST
A)	Decollate SDS Data Forms	$ 3,000
B)	SDS Opscanning	2,000
C)	Edit SDS Data	5,000
D)	Set-up and Load SDS Files	6,000
E)	Decollate CDS Data Forms	2,000
F)	Keypunch CDS Data	40,000
G)	Set-up and Load CDS Files	6,000
H)	Check and Edit CDS Data	3,000
I)	Operate System	28,000
J)	Terminal Rental	20,000
	TOTAL	$115,000